Are residues in a protein folding nucleus evolutionarily conserved?

Yan Yuan Tseng and Jie Liang[*]
Department of Bioengineering, SEO, MC-063
University of Illinois at Chicago
851 S. Morgan Street, Room 218
Chicago, IL 60607–7052, U.S.A.
(Accepted by *J. Mol. Biol.*)

February 9, 2008

Abstract

Protein is the working molecule of cell, and evolution is the hallmark of life. It is important to understand how protein folding and evolution influences each other. Several studies correlating experimental measurement of residue participation in folding nucleus and sequence conservation have reached different conclusions. These studies are based on assessment of sequence conservation at folding nucleus sites using entropy or relative entropy measurement derived from multiple sequence alignment. Here we report analysis of conservation of folding nucleus using an evolutionary model alternative to entropy based approaches. We employ a continuous time Markov model of codon substitution to distinguish mutation fixed by evolution and mutation fixed by chance. This model takes into account bias in codon frequency, bias favoring transition over transversion, as well as explicit phylogenetic information. We measure selection pressure using the ratio ω of synonymous *vs.* non-synonymous substitution at individual residue site. The ω-values are estimated using the PAML method, a maximum-likelihood estimator. Our results show that there is little correlation between the extent of kinetic participation in protein folding nucleus as measured by experimental ϕ-value and selection pressure as measured by ω-value. In addition, two randomization tests failed to show that folding nucleus residues are significantly more conserved than the whole protein, or the median ω value of all residues in the protein. These results suggest that at the level of codon substitution, there is no indication that folding nucleus residues are significantly more conserved than other residues. We further reconstruct candidate ancestral residues of the folding nucleus and suggest possible test tube mutation studies for testing folding behavior of ancient folding nucleus.

[*]Corresponding author. Phone: (312)355–1789, fax: (312)996–5921, email: `jliang@uic.edu`

1

Are amino acid residues important for rapid folding preferentially conserved during evolution? Does natural selection optimize proteins for folding kinetics? If protein folding involves initially the formation of a small region of native-like folding nucleus, are identities of these residues well conserved during evolution [1–5]? These fundamental questions of molecular biology have received much attention [1–12]. Of direct relevance are experimental ϕ-value studies, which provide information about the role individual amino acid residue play in the formation of folding nucleus [13–15]. By measuring the change $\Delta\Delta G$ in protein stability and the change $\Delta\Delta G^{\ddagger}$ in folding barrier due to mutation of an amino acid residue, ϕ-value (defined as $\phi \equiv \Delta\Delta G^{\ddagger}/\Delta\Delta G$) for the mutated residue can be calculated. ϕ-value has been used to measure the extent to which side chain of a mutated residue participates in native-like interactions. A ϕ-value of 0.0 indicates that the site of mutation is as unfolded as in the denatured state. A ϕ-value of 1.0 indicates that the site of mutation is as folded as in the native state, $i.e.$, this residue is involved in native-like transition state structure, and is part of the folding nucleus. ϕ-value between 0 and 1 is interpreted as possessing different degrees of structure in transition state [13]. The folding nucleus can be identified as formed by the set of residues with ϕ values above a threshold ($e.g.$, $\phi \geq 0.5$) [13]. Several computational methods have been developed for predicting protein folding mechanism and ϕ-values of residues. These include the sequential binary collision model [16], multisegment model [17], and single-to-triple sequence approximation model [18]. Model conformations of transition-state ensemble have also been generated explicitly by Monte Carlo sampling using Gō-type potential derived from experimental ϕ-values constraints [19]. A lucid statistical mechanistic picture for understanding ϕ-value experiments can be found in [20, 21].

The evolutionary conservation of folding nucleus residues are the subject of several recent studies. These studies, however, have come to different conclusions. Plaxco et al and Larson et al showed that there may be little correlation between sequence conservation and participation in the folding transition state [8, 9]. Mirny, Shahknovich and others demonstrated that for rapid folding, sequence identity of folding nucleus are more conserved within protein families and across protein superfamilies [2, 7]. It is unclear whether the disagreement between these studies is due to the difference in entropy calculations as attributed in [7], or differences in choice and processing of the data set, in sequence alignments, in definition of folding nucleus, as well as intrinsic sample bias in ϕ-value analysis, as discussed in details in [9].

In this study, we examine the conservation of folding nucleus residues using an approach that differs from previous studies in several aspects. First, instead of studying amino acid residue sequences, we examine the evolution of corresponding coding DNA sequences at codon level. Second, we use an explicit codon evolutionary model based on continuous time Markov process, which has yielded deep insights about the mechanisms of molecular evolution [22–24]. Instead of using entropy or relative entropy as quantitative measure of sequence conservation, we assess the ratio of mutation rates of synonymous $vs.$ non-synonymous changes to detect natural selection at each amino acid residue position. Third, a phylogenetic tree is built to encode the closeness between proteins. Following previous studies [25, 26], we use maximum likelihood method developed in [27] by Yang to estimate values of parameters of the evolutionary model and draw inference about the conservation of folding nucleus residues.

We find experimental ϕ-values are not correlated with evolutionary conservation for seven proteins studied here. In addition, results using two statistical tests indicate that except possibly one protein, none of these proteins have folding nucleus more conserved than the rest of the proteins, or than the residue with median selection pressure. We have also reconstructed candidate ancestral folding nucleus residues, and have suggested exploratory test-tube mutation studies on the evolution of protein folding dynamics.

Synonymous and nonsynonymous codon substitution. Protein sequences diverge from a common ancestor because mutations occur. Some fraction of these mutations are fixed into the evolving population by selection and some are fixed by chance, resulting in the substitution of one nucleotide for another nucleotide at various locations. Because evolution occurs at DNA level rather than at amino acid level, models of protein evolution based on codon usage are appealing and have been widely used [25, 28–30]. In this study, we therefore consider substitutions at the codon level. A codon substitution can have two different outcomes for the nucleotide sequence of protein coding region: $synonymous$

substitution does not change the encoded sequence of amino acid residues, whereas *nonsynonymous substitution* leads to changes in the amino acid residues. Random mutation and selection pressure will have different effects on the rate of these two types of substitutions [31–33], and this difference can be exploited for detecting selective pressure at protein level [25, 34–38]. Our key problem is to find out the ratio of the synonymous substitution rate d_s and the nonsynonymous substitution rate d_n. That is, we wish to estimate the ratio of the numbers of synonymous and non-synonymous substitutions at a specific site or a specific amino acid residue position. If natural selection offers no advantage, non-synonymous mutations will have the same rate as synonymous mutations ($d_n = d_s$), and the ratio $\omega = d_n/d_s$ will be 1. If non-synonymous mutations are harmful, deleterious or lethal, purifying selection is at play and the rate for non-synonymous mutation will be reduced: we have $d_n < d_s$ and $\omega < 1$. On the other hand, if Darwinian positive selection favors non-synonymous mutation, we have $d_n > d_s$ and $\omega > 1$. Here ω is used as a measure of selection pressure. Substitution fixed by evolution and substitution fixed by chance are distinguished by examining the ratio ω at various locations of amino acid residues. This technique has been frequently applied in studies of molecular evolution, *e.g.*, in detecting adaptive evolution [22, 38, 39].

Continuous time Markov process for codon substitution. Markov model has been widely used in sequence analysis [40] and in evolutionary models [23]. In the current model, the outcome of codon substitution is determined only by the identity of codon in the ancestral sequence separated by divergence time t, and a codon transition probability matrix $\mathbf{P}(t)$. A phylogenetic tree is a key ingredient of this model. The topology and branch lengths of the tree reflects the evolutionary relationship among different proteins, which can model their closeness [23]. We follow the approach of [22, 26, 41], and describe below briefly the model.

For a given phylogenetic tree, the parameters of the evolutionary model are a 61×61 rate matrix $\mathbf{Q}$ for 61 non-stop codons and the sequence divergence time ts (or the branch lengths) of the phylogenetic tree. The divergence time represents expected number of changes between sequences which are nodes in a phylogenetic tree. The entries q_{ij} of matrix $\mathbf{Q}$ are infinitesimal substitution rates of nucleotides for the set $\mathcal{C}$ of 61 non-stop codons, and they are parametrized as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion,} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition,} \end{cases}$$

where μ is the basis rate, κ is the transition/transversion rate ratio, ω the ratio of nonsynonymous and synonymous rates, and π_j is the codon frequency, which can be estimated as observed codon frequency in the sequences. In this model, the 61×61 rate matrix $\mathbf{Q}$ is fully determined by two parameters κ and ω, since π_j can be estimated and μ is a constant [25, 26].

For continuous time Markov process, the transition probability matrix of size 61×61 after time t is [24]:

$$\boldsymbol{P}(t) = \{p_{ij}(t)\} = \exp(\boldsymbol{Q} \cdot t)$$

The entry $p_{ij}(t)$ represents the probability that codon i will mutate into codon j after time t. It is calculated through diagonalization of the $\boldsymbol{Q}$ matrix.

ω **ratio from likelihood of phylogeny.** For node i and node j in a phylogenetic tree separated by divergence time t_{ij}, the time reversible probability of observing nucleotide x_i in a position h at node i and nucleotide x_j of the same position at node j is:

$$\pi_{x_i} p_{x_i x_j}(t_{ij}) = \pi_{x_j} p_{x_j x_i}(t_{ij}). \tag{1}$$

For a set $\mathcal{S}$ of s multiple-aligned sequences with n amino acid residues, we assume that a reasonably accurate phylogenetic tree $\boldsymbol{T} = (\mathcal{V}, \mathcal{E})$ is given. Here $\mathcal{V}$ is the set of nodes (or vertices), namely, the union of the set of observed s sequences $\mathcal{L}$ (leaf nodes), and the set of $s - 2$ ancestral sequences $\mathcal{I}$ (internal nodes). $\mathcal{E}$ is the set of edges (or branches) of the tree. Let the vector $\boldsymbol{x}_h = (x_1, \cdots, x_s)^T$ be the observed

codons at position h for the s sequences. Without loss of generality, we assume that the root of the phylogenetic tree is an internal node k. Given the specified topology of the phylogenetic tree $\mathbf{T}$ and the set of branch lengths (or divergence times), and if the set of codons $\mathcal{C}_\mathcal{I}$ of all internal nodes $\mathcal{I}$ is specified, the probability of observing the s number of codons $\boldsymbol{x}_h$ at position h is:

$$p(\boldsymbol{x}_h|\mathcal{C}_\mathcal{I}, \boldsymbol{T}) = \pi_{x_k} \prod_{(i,j)\in\mathcal{E}} p_{x_i x_j}(t_{ij}).$$

Summing over the set $\mathcal{C}$ of all possible codons for the internal nodes $\mathcal{I}$, we have

$$p(\boldsymbol{x}_h|\boldsymbol{T}) = \pi_k \sum_{\substack{i\in\mathcal{I} \\ x_i\in\mathcal{C}}} \prod_{(i,j)\in\mathcal{E}} p_{x_i x_j}(t_{ij}). \tag{2}$$

The probability of observing all codons in the coding region of the nucleotide sequences is:

$$P(\mathcal{S}|\boldsymbol{T}) = P(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s|\boldsymbol{T}) = \prod_{h=1}^{s} p(\boldsymbol{x}_h|\boldsymbol{T}).$$

To account for the possibility that the rate of nonsynonymous substitution can vary among different sites, the model developed in [41] allows M possible different classes of nonsynonymous substitutions with rates $\omega_1, \cdots, \omega_M$. Each amino acid site falls into the M class with probabilities $p_1, \cdots, p_M$ [41]. The probability of observing $\boldsymbol{x}_h$ is then modified from Equation (2), which gives $p(\boldsymbol{x}_h|\omega_m, \boldsymbol{T})$, to the following:

$$p(\boldsymbol{x}_h|\boldsymbol{T}) = \sum_{m=1}^{M} p_m \cdot p(\boldsymbol{x}_h|\omega_m, \boldsymbol{T}).$$

Repeat this calculation over all amino acid residue sites, we have

$$P(\mathcal{S}|\boldsymbol{T}) = \prod_{h=1}^{s} p(\boldsymbol{x}_h|\boldsymbol{T}),$$

and the likelihood function is:

$$\ell(\boldsymbol{T})) = \sum_{h=1}^{s} \log[p(\boldsymbol{x}_h|\boldsymbol{T})].$$

To estimate the parameters κ_h, ω_h for each site h used in the mutation rate matrix $\boldsymbol{Q}$, we use a Maximum Likelihood Estimator [26, 37, 42], the PAML package by Yang [27]. Our goal is to search for parameters κ_h and ω_h such that the likelihood function $\ell(\boldsymbol{T})$ is maximized. Here the number M of different classes of ω is 10, and they take the default values as assigned by PAML [41].

Once the model parameters are estimated, the empirical Bayes approach can be used to infer the most likely class of ω value at each residue site [22]. In PAML, the posterior probability $p(\omega_m|\boldsymbol{x}_h)$ that site h with observed codons $\boldsymbol{x}_h$ is from class m with rate ratio ω_m is calculated as:

$$p(\omega_m|\boldsymbol{x}_h) = p_m \cdot p(\boldsymbol{x}_h|\omega_m, \boldsymbol{T})/p(\boldsymbol{x}_h|\boldsymbol{T}) = p_m \cdot p(\boldsymbol{x}_h|\omega_m, \boldsymbol{T})/\sum_m p_m \cdot p(\boldsymbol{x}_h|\omega_m, \boldsymbol{T}).$$

Data collection and computational procedures. We follow [7] and study evolution of the set of proteins taken from Table 1 of [7], where the folding nucleus residues are defined. We first query with the sequence of each of the proteins against HSSP database [43] to obtain homologous protein sequences with overall sequence identity $> 30\%$ to ensure that they have the same fold. In some cases, we also searched the CE server [44] for structural homologs. Experimentation using PSI-BLAST searching the NR-database of protein sequences give almost identical sets of sequences. In this study, all redundant sequences are removed. Since paralogous sequences in a single species may exist that can be matched to the query DNA sequence, we only take the sequence with the highest identity to the query protein when
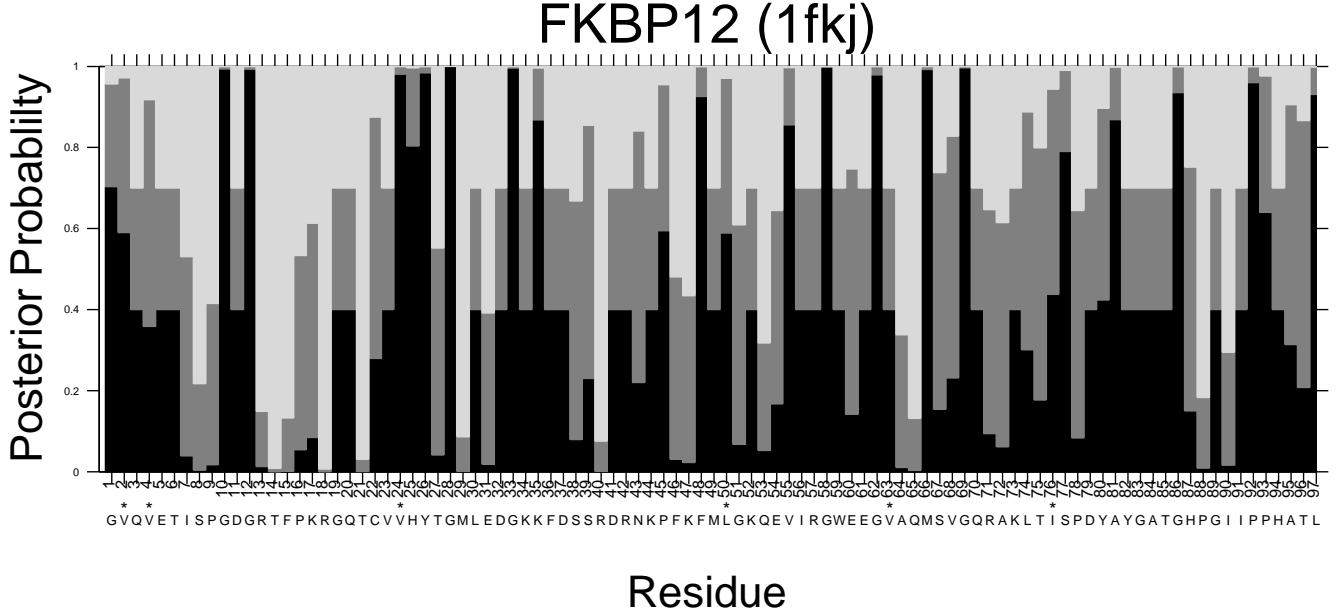
FKBP12 (1fkj)

Figure 1: Selection pressure as measured by ω ratio of nonsynonymous *vs.* synonymous codon substitution rate varies at each amino acid residue site along the sequence of protein FKBP12. The ten possible ω values are grouped into three classes: $\omega_a < 0.12$ (dark), $0.12 \leq \omega_b < 0.34$ (gray), and $0.34 < \omega_c$ (light). The x-axis shows the residue number of the protein, the y-axis shows the posterior probability of ω belonging to one of the three classes at each codon position. Residues with large probability for ω_a (dark) are highly conserved residues experiencing strong purifying pressure. Folding nucleus residues as identified in [7] are marked by the symbol "*".

multiple homologous sequences are found in a single species. With the exception of protein CI2 where sequences of two paralogs are included, only proteins with ≥ 5 known orthologous DNA sequences are kept. We therefore exclude AcP protein and CD2.d1 protein because fewer than 5 DNA sequences were found. Since paralogs are excluded, the number of sequences used in this study is smaller than that used in other studies [7–9]. The amino acid residue sequences of the remaining 7 proteins are first aligned using CLUSTALW with default parameters [45] and then with manual intervention. Alignment of the nucleotide sequences are generated following the alignment of the protein sequences. A phylogenetic tree **T** is constructed using maximum likelihood method as implemented in the PAUP method [46]. This tree **T** is then used by the PAML package, an implementation of the maximum likelihood method for estimating ω values [27]. In many cases, minor difference in the tree does not affect final results significantly [47, 48]. For each protein, we repeatedly estimate ω twenty times using different initial ω value that is assigned to all amino acid sites. The initial ω values range from 0.01 to 2.00, at an interval of 0.1. About 90% of the computation converges. For each protein, all different converged estimations among the twenty calculations give identical ω parameters at individual codon positions.

Natural selection at protein folding nucleus. The estimation of site-specific ω-values can uncover residues important for biological function, for structural stability, and potentially for folding kinetics. In this study we focus on the natural selection of folding nucleus residues which are identified by ϕ-value experiments. An example for estimated ω values is shown in Fig 1.

We first examine the patterns of ω-ratio of nonsynonymous *vs.* synonymous substitutions in the seven proteins. If folding nucleus residues are more conserved than other residues, selection pressure then must be correlated with the extent of participation in folding nucleus [9]. Following Larson *et al*, we examine directly the correlation of the ϕ-values and the ω-values of characterized residues for each protein. This approach helps to circumvent the unavoidable arbitrariness in the assignment of the set
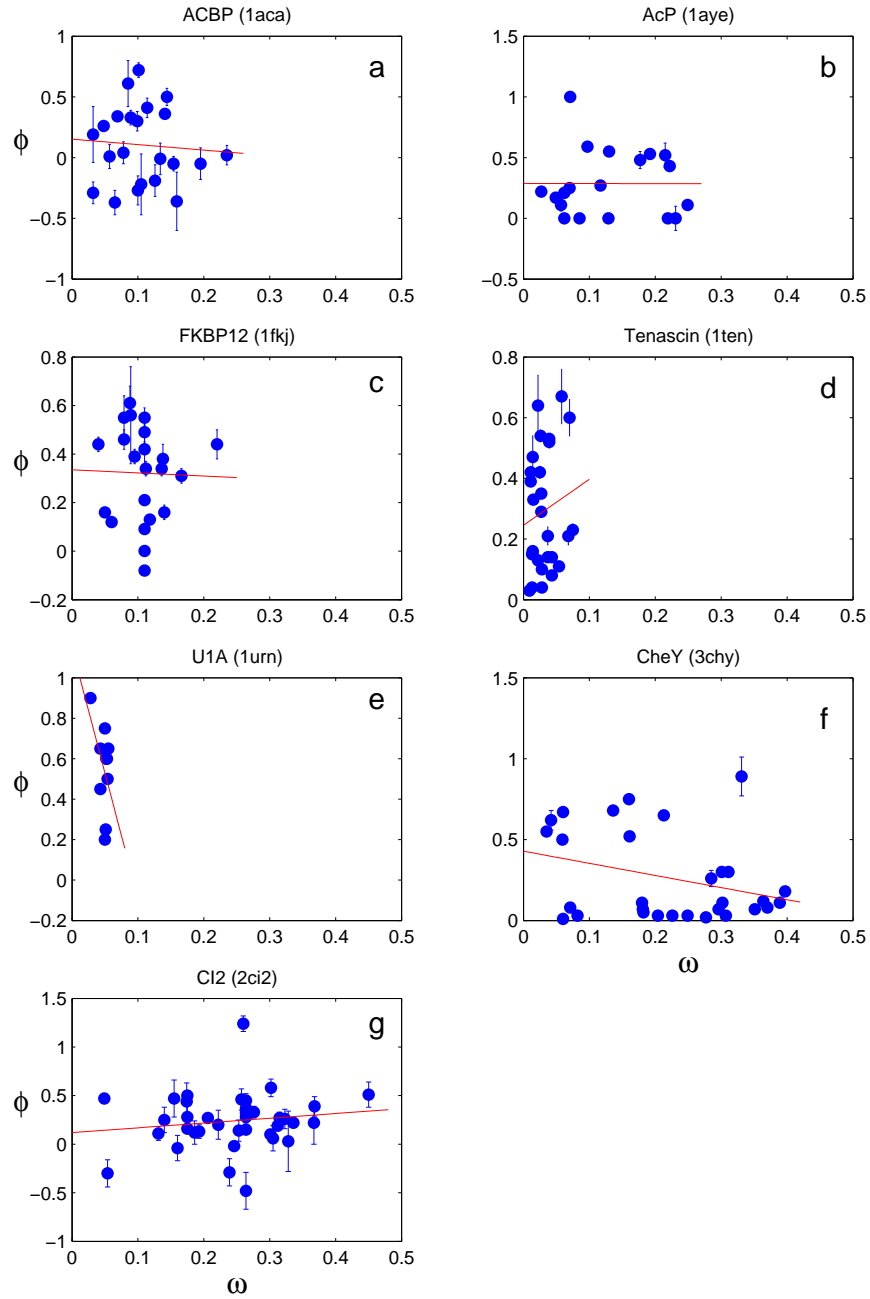
Figure 2: Participation in folding nucleus as measured by experimental ϕ-value and selective pressure as measured by ω-value are poorly correlated.

of folding nucleus residues [9, 13]. Residues with characterized ϕ-values for these proteins are obtained from references cited in [9]. Following [8], we exclude residues with $\phi < -0.5$ or $\phi > 1.5$, and require all ϕ-values to have standard deviation < 1.0, with the exception of protein U1A (1urn), where no data of standard deviations are provided.

Among the set of residues with experimentally characterized ϕ-values, there is little correlation between ϕ-value and ω-value (Fig 2). The R^2 values range between 0.0 to 0.22, and the two-sided
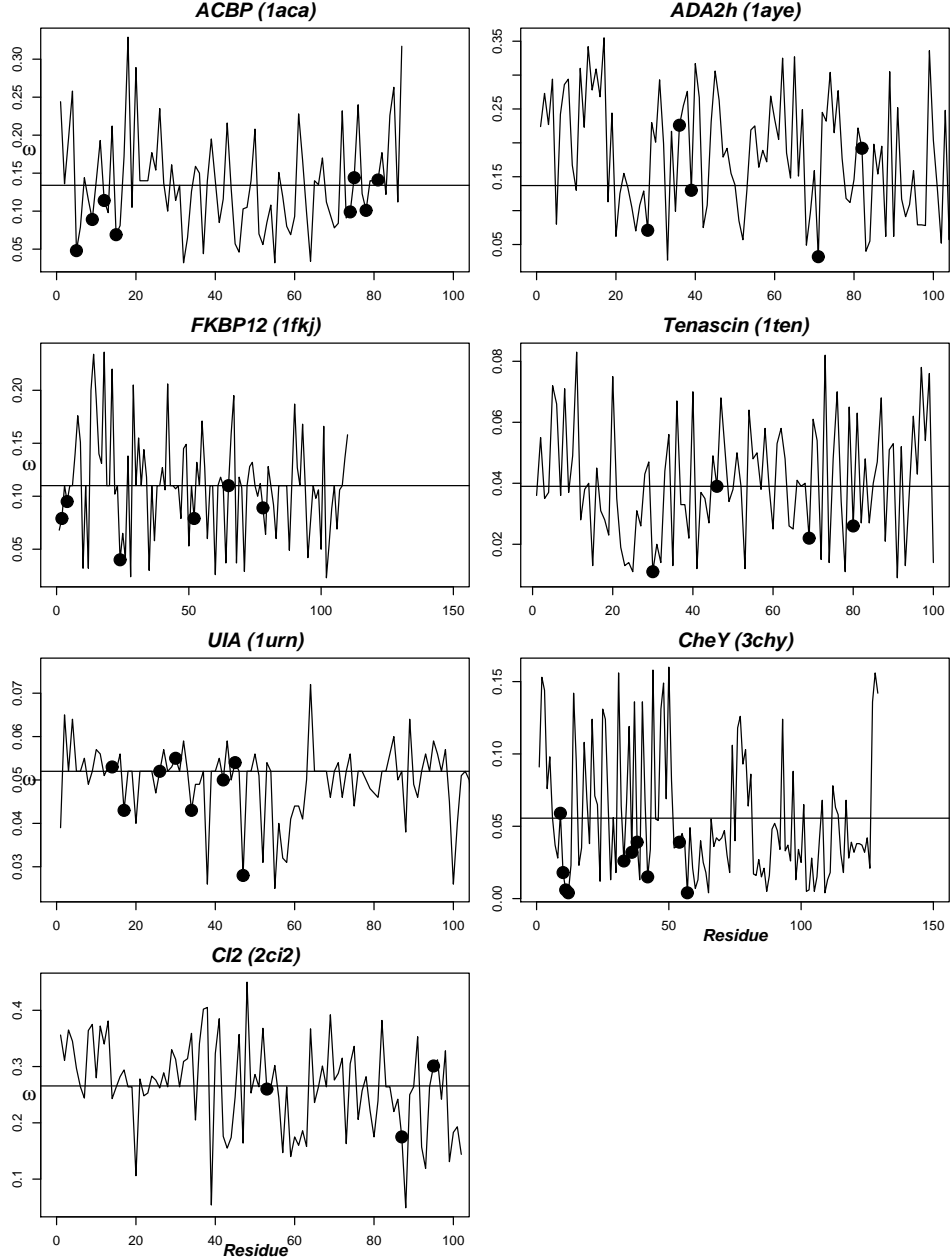
Figure 3: The weighted mean value $\bar{\omega} = \sum_{m=1}^{10} p_m \cdot \omega_m$ of estimated ω ratio at each residue position of the proteins. The x-axis shows the residue number of the protein, the y-axis shows the estimated $\bar{\omega}$ at each residue position. The horizontal line marks the median $\bar{\omega}$ value of all positions. Folding nucleus residues as identified in [7] are marked by "*". Except protein CheY, randomization tests show that folding nucleus residues are not more conserved than the rest of the protein, and in all protein cases (including CheY protein) are not more conserved than the residue at 50% quantile of all residues ranked by ω.

p-values of t-test for the null hypothesis that the slope of the linear regression models is 0 range from 9% to 99% (Table I). That is, there is no indication of significant correlation between the extent of kinetic participation as measured by ϕ-value and selection pressure as measured by ω-value. Our results are similar to those found in [8, 9], where relative entropy instead of ω was used as the measure of

7

Protein	PDB	N_{prot}	N_{seq}	N_ϕ	$^a R^2$	$^a p$	$^b p_{\text{all}}$	$^b p_{50\%}$	$^c Z_{\alpha,\text{fn}}$	$^c Z_{\alpha,\text{all}}$
CI2	2ci2I	83	5	37	2.1×10^{-2}	0.39	3.3×10^{-1}	8.6×10^{-1}	3.29	2.81
Tenascin	1ten	2201	5	27	2.1×10^{-2}	0.47	4.1×10^{-2}	2.5×10^{-1}	3.29	3.44
CheY	3chy	128	7	30	9.8×10^{-2}	0.093	2.6×10^{-3}	8.2×10^{-2}	3.60	3.25
ADA2h	1aye	417	6	19	5.9×10^{-6}	0.99	4.3×10^{-1}	9.9×10^{-1}	3.43	2.78
U1A	1urn	282	12	10	2.2×10^{-1}	0.17	1.6×10^{-1}	9.3×10^{-1}	3.35	3.48
ACBP	1aca	86	16	22	5.2×10^{-3}	0.75	6.7×10^{-2}	6.3×10^{-1}	NMR	NMR
FKBP12	1fkj	107	27	22	6.7×10^{-4}	0.91	4.0×10^{-2}	3.6×10^{-1}	3.11	2.99

Table I: The conservation and packing of folding nucleus residues. N_{prot}: number of residues in the protein sequence; N_{seq}: number of sequences; N_ϕ: number of residues with ϕ-value measured. (a) Correlation of participation in folding nucleus as measured by ϕ-value and selection pressure as measured by ω. R^2: the fraction of variance in the data that can be explained by the linear regression model; p: the two-sided p-value of t-test for the null hypothesis that the slope of the linear regression models is 0. (b) Randomization tests for assessing statistical significance of conservation of folding nucleus residues. The median ω value of the folding nucleus is tested against the distribution of the median ω value from 10^5 random samples containing the same number of amino acid residues as that of the folding nucleus drawn from the same protein. p_{all}: the p-value that the folding nucleus residues are more conserved than all other residues in the protein; $p_{50\%}$: the p-value that folding nucleus residues are more conserved than the residue at 50% quantile of all residues ranked by ω-value. (c) Packing analysis of the folding nucleus and of the whole protein. The average alpha coordination number Z_α for all residues in the protein ($Z_{\alpha,\text{all}}$) and for residues in the folding nucleus residues ($Z_{\alpha,\text{fn}}$) are listed, except for structures determined by NMR techniques. Protein CheY has the highest $Z_{\alpha,\text{fn}}$.

evolutionary conservation.

The weighted mean values of estimated ω ratio $\bar{\omega} = \sum_m p_m \cdot \omega_m$ at each codon position are plotted in Figure 3. It is clear that for each protein, many folding nucleus residues as defined in [7] have small values of ω, many are often smaller than the median ω-value of all codon positions. This indicates that folding nucleus residues experience purifying selection pressure. However, there are also many other residues with small ω-value, some of which have not be characterized by ϕ-value studies. As discussed in [9], the lower ω-values of folding nucleus as defined in [7] residues could also be a reflection of the experimental bias in choosing conserved protein core residues for ϕ-value experiments. Can we still conclude that experimentally identified folding nucleus residues in general are more conserved than the rest of the protein?

We use a randomization test following the approach first developed in [7] to address this question. The null hypothesis H_0 is that nucleus residues have equal or greater median ω values than that of the whole protein. That is, folding nucleus residues are no more conserved than the whole protein sequence. The alternative hypothesis H_a is that folding nucleus residues have less median ω values than the whole protein sequence and are evolutionarily more conserved. We calculate the median of ω values of the nucleus residues as defined in [7], and compare it with the distribution of median of ω value in random samples containing the same number of residues drawn from the same protein. As in [7], we use a sample size of 10^5. The fraction of the random samples with median ω value smaller than that of the folding nucleus provides the p-value that the observed median ω-values of the folding nucleus is due to random chance. Similar to [7], we use the threshold of $p = 2\%$ to decide whether evolutionary conservation of the folding nucleus is statistically significant. Table I shows that p-value ranges between 0.26% (CheY) and 43% (ADA2h), but the majority are between 4.0% (FKBP12) and 43% (ADA2h). With the exception of CheY, the null hypothesis cannot be rejected with statistical significance at the confidence level of $p < 2\%$. That is, except CheY, folding nuclei as defined in [7] in these proteins are not significantly more conserved than the rest of the protein.

To further assess selection pressure on folding nucleus residues, we evaluate a different null hypothesis,

again using randomization test. The null hypothesis H_0 now is that the folding nucleus residues have equal or greater median ω-values than the residue with median ω-value of the whole protein. That is, folding nucleus as defined in [7] are no more conserved than the residue halfway in the rank ordered list of all residues when sorted by estimated mean ω-value. Table I shows that the p-values range from 8.2% to 99%. With the criterion of $p < 2\%$, the null hypothesis cannot be rejected with statistical significance for any of the proteins. That is, folding nucleus for every protein studied here is not significantly more conserved than the residue with median ω-value.

Conservation of folding nucleus of CheY. CheY is the only protein among those studied here that may have a well-conserved folding nucleus based on results of the first readomization test. Correlation study of ϕ-value and conservation measured by reduced entropy also suggested that CheY protein has a well-conserved folding nucleus [9]. What are the possible reasons for the strong conservation of folding nucleus in this protein? It was suggested earlier that tightly packed protein interior residues are well conserved and these are often part of the folding nucleus residues [4, 6, 49]. We use a parameter z_α recently introduced in [50] to characterize protein local packing. z_α is defined as $z_\alpha \equiv n_c/n$, where n_c is the number of non-bonding atomic alpha contacts between different residues, and n is the total number of atoms. Two atoms are in alpha contact if they are separated by a weighted Voronoi facet which intersects with the protein [50]. z_α characterizes protein packing more faithfully than other parameters such as radius of gyration [50].

We calculate z_α for the folding nucleus as defined in [7] and for the whole protein (Table I). We find that the folding nucleus of CheY has the highest z_α value (3.60) compared to the folding nuclei of other proteins, whereas the whole protein z_α value of CheY has similar values to other proteins. This indicates that the folding nucleus of CheY has significantly larger z_α than the rest of CheY protein. The folding nucleus of CheY is packed tighter than folding nuclei in other proteins. This observation can intuitively explain the significant conservation in CheY: tight packing in this case is accompanied by little tolerance to mutation, since the lack of packing defects such as voids reduces the possibility for substitution of different amino acid residues. However, this is a rather tentative hypothesis. It is possible that very tightly packed residues are more conserved, independent of whether they are in folding nucleus or not. It is also possible that if results of additional experimental ϕ-value studies become available, the definition of the folding nucleus might change. To fully resolve the relationship of packing, folding, and evolutionary conservation, more detailed additional studies are required, which is beyond the scope of this work.

Reconstructing ancestral folding nucleus. The approach used in this study can also suggest further experimental exploration of evolution history of protein folding dynamics. With the continuous time Markovian model, we can reconstruct likely candidate sequences of ancestral proteins at different evolutionary times. Specifically, identities of amino acid residues in the folding nucleus of ancient ancestral proteins can be postulated.

As an example, we show the reconstructed residues of the folding nuclei of FKBP12 as defined in [7] in Figure 4. The five folding nucleus residues are VVVLVI in human FKBP12 protein. The first residue is L in some reconstructed ancestral genes, the second can be Y or N, the third can be L, and the sixth can be a V instead of I. Based on this simple analysis, an interesting quadruplet mutagenesis study can be suggested to experimentally test the folding dynamics of mutated FKBP12, where the folding nucleus is changed. The reconstructed ancient folding nuclei suggests a combination of residues represented by the pattern **L{N,Y}L***LV***V**. Here **{N,Y}** means either a N or a Y residue is drawn.

The fourth residue L and fifth residue Y in all ancestral genes are the same as that in human FKBP12, but inspection of sequences of other extant species show that the fourth residue can be any of I, P, or V, and the fifth can be any of I, V, L, and M. A further interesting experiment could be to test the folding behavior of 6-tuple mutants with folding nucleus formed by any combination of residues represented by the pattern **L{N,Y}{I,P,V}{I,V,L,M}V**. The recreated proteins then can be assayed for folding behavior, which can be compared with that of proteins present in extant organisms. Such experimental palaeobiochemistry was already envisioned by Pauling and Zuckerkandl many years ago [51], and the number of such studies is rapidly growing [52–57]. An in-depth study on recreating the full sequence of ancestral proteins will require additional detailed analysis, including choosing the most appropriate
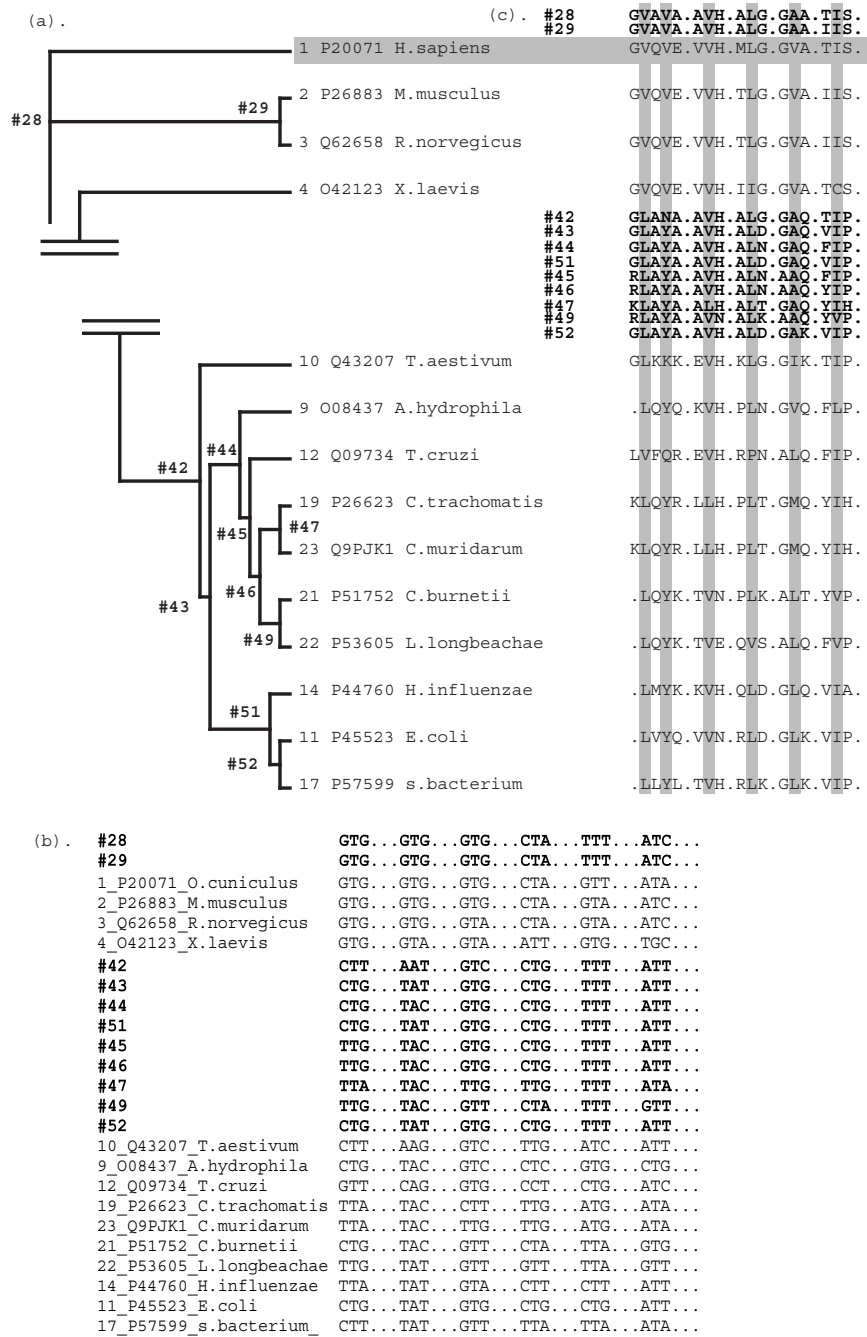
Figure 4: Reconstructed ancestral protein sequences of FKBP12 protein. (a). The relevant part of the phylogenetic tree for FKBP12 is shown. Human FKBP12 protein from which experimental data were obtained is shown in shadow. (b). Multiple alignment of DNA sequences of the folding nucleus of FKBP12 protein, including those of reconstructed folding nucleus of ancestral proteins. (c). Multiple alignment of translated amino acid residue of the folding nucleus residues identified by ϕ-value studies (highlighted) and flanking residues.

detailed evolutionary model [58–60].

Discussion. Although folding nucleus is under purifying pressure, we fail to observe significant

conservation for protein folding nucleus residues. Despite concerns raised in [9] about the specific choices of the data in [7], we use exactly the same set of proteins, the same definition of nuclei residues, and follow the same radomizatin test as that of [7]. It is possible that this would bias our study towards reproducing the results of [7]. Nevertheless, our results are similar to that of Plaxco *et al* and Larson *et al* [8, 9], and are different from that of Mirny and Shakhnovich [7]. The different conclusion of this study and that of [7] is likely due to the different evolutionary models employed, namely, the difference between a DNA-codon based continuous-time Markov model *vs.* an implicit evolution model implied by entropy calculation. The conclusion that folding nuclei residues are not conserved will likely to remain if we were to use the data set and the definitions of folding nuclei from reference [9]. Experimental studies in barnase, SH3 domain, chymotrypsin inhibitor 2 suggest that the folding nucleus observed in wild type protein may not be indispensable, and alternative folding nucleus may arise if residues are mutated [61–65]. Another experimental example is Im9 and Im7 proteins. They are E colicin-binding immunity proteins that are of the same fold with about 60% sequence identity. The folding of Im9 and Im7 are two-state and three-state process, respectively. Although these two proteins have similar folding mechanism, ϕ-value studies reveal that the kinetically important residues are different [66, 67]. This is consistent with recent simulation studies which suggest that evolution selection is more robust for residues important for stability than for kinetic accessibility [68, 69]. In addition, the definition of a folding nucleus is arbitrary, because it is based on a threshold of ϕ value (*e.g.*, $\phi \geq 0.5$) [13]. An earlier study suggested that the critical nucleus may be as large as 10^2 residues, the size of a whole protein domain [70]. The non-uniqueness of folding nucleus was pointed out in a study using off-lattice model system [71]. The role of protein structure in folding is discussed from the viewpoint of small-world connections in [72]. Recent computational studies based on exact enumerable lattice models using master equation showed that there are remarkable heterogeneity in structural contacts underlying macroscopic two-state folding kinetics of model Gō protein [20, 21]. The kinetic barrier was shown to result from a reduced number of microroutes near the bottom of the folding funnel [20, 21]. If these studies portray accurately the microscopic picture of the folding process, there are likely to be many different native contacts that form folding nuclei for different folding pathways in the free energy landscape. It is reasonable to expect that a large subset of residues are capable of providing critical native contacts, and these contacts vary for different microscopic folding pathways. The roles of these residues in folding are largely interchangeable, and this may be reflected in the lack of extraordinarily strong purifying selection pressure in the current set of folding nucleus residues characterized by ϕ-value studies.

In summary, we use a continuous time Markovian model [25] and apply a maximum likelihood estimator developed in [27] to study the evolution of protein folding dynamics. We examine the coding DNA sequences rather than amino acid residue sequences, and assess selection pressure by estimating the ratio ω of nonsynonymous *vs.* synonymous codon substitution rate. The position specific rate ratio is used to distinguish substitutions fixed by evolution and by chance. We found that folding nucleus residues experience purifying selection pressure, but they are not significantly more conserved than the rest of the residues of the whole protein. The only exception is CheY protein, where the folding nucleus is significantly more conserved. This may be due to extraordinarily tight packing, which is reflected by the high alpha coordination number Z_α. Results described here provides another confirmation that evolution does not preserve kinetically important residues, which has been a subject of debate in literature [7–9]. We further suggest exploratory palaeobiochemical studies testing the evolution of protein folding dynamics.

References

[1] I. Shrivastava, S. Vishveshwara, M. Cieplak, A. Maritan, and J.R. Banavar. Lattice model for rapidly folding protein-like heteropolymers. *Proc Natl Acad Sci USA*, 92:9206–9209, 1995.

[2] E. Shakhnovich, V. Abkevich, and O. Ptitsyn. Conserved residues and the mechanism of potein folding. *Nature*, 379:96–98, 1996.

[3] L.A. Mirny, V.I. Abkevish, and E.I. Shakhnovich. How evolution makes proteins fold quickly. *Proc Natl Acad Sci USA*, 95:4976–4981, 1998.

[4] O.B. Ptitsyn. Protein folding and protein evolution: Common folding nucleus in different subfamilies of *c*-type cytochromes. *J. Mol. Biol.*, 278:655–666, 1998.

[5] S.W. Michnick and E. Shakhnovich. A strategy for detecting the conservation of folding -nucleus residues in in protein superfamilies. *Fold. Des.*, 3:239–251, 1998.

[6] O.B. Ptitsyn and K-L. H. Ting. Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.*, 291:671–682, 1999.

[7] L. Mirny and E. Shakhnovich. Evolutionary conservation of the folding nucleus. *J. Mol. Biol.*, 308:123–129, 2001.

[8] K.W. Plaxco, D.S. Riddle, S. Larson, I. Ruczinski, E.C. Thayer, B. Buchwitz, A.R. Davidson, and D. Baker. Evolutionary conservation and protein folding kinetics. *J. Mol. Biol.*, 298:303–312, 2000.

[9] S.M. Larson, I. Ruczinski, A.R. Davidson, D. Baker, and K.W. Plaxo. Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.*, 316:225–233, 2002.

[10] V.P. Grantcharova, D.S. Riddle, J.V. Santiago, and D. Baker. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct Biol*, 5:714–720, 1998.

[11] K.F. Fulton, E.R.G. Main, V. Dagget, and S.E. Jackson. Mapping the interactons present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.*, 291:445–461, 2000.

[12] M.C. Demirel, A.R. Atilgan, R.L. Jernigan, B. Erman, and I. Bahar. Identification of kinetically hot residues in proteins. *Protein Sci*, 7:2522–2532, 1998.

[13] A.R. Fersht. Nucleation mechanism in protein folding. *Current Opinion in Structural Biology*, 7:3–9, 1997.

[14] A. Matouschek, J.T. Kellis Jr, L. Serrano, and A.R. Fersht. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, 346:440–445, 1990.

[15] A. Matouschek and A.R. Fersht. Protein engineering in analysis of protein folding pathways and stability. *Methods Enzymol.*, 202:82–112, 1991.

[16] E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc Natl Acad Sci U S A*, 96:11305–11310, 1999.

[17] O.V. Galzitskaya and A.V. Finkelstein. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc Natl Acad Sci U S A*, 96:11299–11304, 1999.

[18] V. Munoz and W.A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci U S A*, 96:11311–11316, 1999.

[19] M. Vendruscolo, E. Paci, C.M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409:641–645, 2001.

[20] S.B. Ozkan, I. Bahar, and K.A. Dill. Transition states and the meaning of ϕ-values in protein folding kinetics. *Nature Structural Biology*, 8:765–769, 2001.

[21] S.B. Ozkan, K.A. Dill, and I. Bahar. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.*, 11:1958–1970, 2002.

[22] Z. Yang. *Handbook of statistical genetics*, chapter 12. John Wiley & Sons, 2001.

[23] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. *Molecular Systematics*. Sinauer, Sunderland, MA, 1996.

[24] P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res.*, 8:1223–1244, 1998.

[25] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evo.*, 11:725–736, 1994.

[26] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amon acid sites and applicatons to the HIV-1 envelope gene. *Genetics*, 148:929–936, 1998.

[27] Z. Yang. PAML: a program for package for phylogenetic analysis by maximum likelihood. *CABIOS*, 15:555–556, 1997.

[28] M. Schöniger, G.L. Hofacker, and B. Borstnik. Stochastic traits of molecular evolution – acceptance of point mutations in native actin genes. *J. Theoretical Biology*, 143:287–306, 1990.

[29] S.V. Muse and B.S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.*, 13:105–114, 1994.

[30] Yang Z., R. Nielsen, and M. Hasegawa. Models of amino acid substitutions and applications to mitochondrial protein evolution. *Mol. Biol. Evol.*, 15:1600–1611, 1998.

[31] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[32] J.H. Gillespie. *The causes of molecular evolution*. Oxford University Press, 1994.

[33] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 11:715–724, 1986.

[34] A.L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335:167–170, 1988.

[35] W.H. Li. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, 36:96–99, 1993.

[36] W. Messier and C.B. Stewart. Episodic adaptive evolution of primate lysozymes. *Nature*, 385:151–154, 1997.

[37] Z. Yang. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.*, 15:568–573, 1998.

[38] Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.*, 46:409–418, 1998.

[39] W.J. Swanson, Z. Yang, M.F. Wolfner, and C.F. Aquadro. Positive darwinian selection in the evolution of mammalian female reproductive proteins. *Proc. Natl. Acad. Sci. USA*, 98:2509–2514, 2000.

[40] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[41] Z. Yang, R. Nielsen, N. Goldman, and A-M.K. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449, 2000.

[42] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[43] C. Sander and R. Schneider. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–58, 1991.

[44] I.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, 11:739–747, 1998.

[45] J.D. Thompson, D.G. Higgins, and T.J Gibson. ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[46] D.L. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony and Other Methods*. Sinauer Associates, Sunderland, MA, 2002.

[47] M.J. Ford. Molecular evolution of transferrin: Evidence for positive selection in salmonids. *Molecular Biology and Evolution*, 18:639–647, 2001.

[48] Z. Yang and W.J. Swanson. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution*, 19:49–57, 2002.

[49] P.L. Privalov. Intermediate states in protein folding. *J. Mol. Biol.*, 258:707–725, 1996.

[50] J. Zhang, R. Chen, C. Tang, and J. Liang. Origin of scaling behavior of protein packing density: A sequential Monte Carlo study of compact long chain polymers. *J. Chem. Phys.*, 118:6102–6109, 2003.

[51] L. Pauling and E. Zuckerkandl. Chemical paleogenetics: molecular 'restoration studies' of extinct forms of life. *Acta Chemica Scandinavica*, 17:S9–S16, 1963.

[52] G.B. Golding and A.M. Dean. The structural basis of molecular adaptation. *Mol. Biol. Evol.*, 15:355–369, 1998.

[53] B.S. Chang and M.J. Donoghue. Recreating ancestral proteins. *Trends in Ecology and Evolution*, 15:109–114, 2000.

[54] N.B. Adey, T.O. Tollefsbol, A.B. Sparks, M.H. Edgell, and C.A. Hutchison III. Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl. Acad. Sci. USA*, 91:1569–1573, 1994.

[55] T.M. Jermann, J.G. Opitz, J. Stackhouse, and S.A. Benner. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, 374:57–59, 1995.

[56] U.M. Chandrasekharan, S. Sanker, M.J. Glynias, S.S. Karnik, and A. Husain. Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science*, 1996.

[57] A.M. Dean and G.B. Golding. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proc. Natl. Acad. Sci. USA*, 94:3104–309, 1997.

[58] D. Schulter. Likelihood of ancestor states in adaptive radiation. *Evolution*, 51:1699–1712, 1997.

[59] C.W. Cunningham, K.E. Omland, and T.H. Oakley. Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.*, 13:361–366, 1998.

[60] J. Zhang and M. Nei. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.*, 44:S139–S146, 1997.

[61] J.M. Matthews and A.R. Fersht. Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry*, 34:6805–6814, 1995.

[62] A.R. Viguera, L. Serrano, and M. Wilmanns. Different folding transition states may result in the same native structure. *Nat. Struct. Biol.*, 3:874–880, 1996.

[63] A.R. Viguera and L. Serrano. Unspecific hydrophobic stabilization of folding transition states. *Proc Natl Acad Sci USA*, 99:5349–5354, 2002.

[64] L.S. Itzhaki, D.E. Otzen, and A.R. Fersht. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, 254:260–288, 1995.

[65] J.L. Neira, B. Davis, A.G. Ladurner, A.M. Buckle, P. Gay Gde, and A.R. Fersht. Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. direct evidence for a nucleation-condensation mechanism. *Fold Des*, 1:189–208, 1996.

[66] A.P. Capaldi, C. Kleanthous, and S.E. Radford. Im7 folding mechanism: misfolding on a path to the native state. *Nat Struct Biol.*, 9:209–216, 2002.

[67] C.T. Friel, A.P. Capaldi, and S.E. Radford. Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol*, 326:293–305, 2003.

[68] N.V. Dokholyan and E.I. Shakhnovich. Understanding hierarchical protein evolution from first principles. *J Mol Biol*, 312:289–307, 2001.

[69] N.V. Dokholyan, L. Li, F. Ding, and E.I. Shakhnovich. Topological determinants of protein folding. *Proc Natl Acad Sci U S A*, 99:8637–8641, 2002.

[70] J.D. Bryngelson and P.G. Wolynes. A simple statistical field-theory of heteropolymer collapse with application to protein folding. *Biopolymers*, 30:1–2, 1990.

[71] Z. Guo and D. Thirumalai. The nucleation-collpse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. *Folding & Design*, 2:377–391, 1996.

[72] M. Vendruscolo, N.V. Dokholyan, E. Paci, and M. Karplus. Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E Stat Nonlin Soft Matter Phys.*, 65(6 Pt 1):061910, 2002.